REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
30-10-2018	Final Report	1-Aug-2015 - 31-Jul-2018

4. TITLE AND SUBTITLE

Final Report: ARO Statistical Foundations for Analyzing Large Collections of Network-Data Objects

5a. CONTRACT NUMBER
W911NF-15-1-0440

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Boston University
Office of Sponsored Program
881 Commonwealth Avenue
Boston, MA 02215 -1300

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
66687-MA.2

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Eric Kolaczyk
UU	UU	UU	UU		19b. TELEPHONE NUMBER 617-353-5208

RPPR Final Report
as of 31-Oct-2018

Agency Code:

Proposal Number: 66687MA **Agreement Number: W911NF-15-1-0440**
INVESTIGATOR(S):

Name: Eric Kolaczyk
Email: kolaczyk@bu.edu
Phone Number: 6173535208
Principal: Y

Organization: **Boston University**
Address: Office of Sponsored Program, Boston, MA 022151300
Country: USA

DUNS Number: 049435266 EIN: 042103547
Report Date: 31-Oct-2018 Date Received: 30-Oct-2018
Final Report for Period Beginning 01-Aug-2015 and Ending 31-Jul-2018
Title: ARO Statistical Foundations for Analyzing Large Collections of Network-Data Objects
Begin Performance Period: 01-Aug-2015 **End Performance Period:** 31-Jul-2018
Report Term: 0-Other

Submitted By: Eric Kolaczyk Email: kolaczyk@bu.edu
Phone: (617) 353-5208

Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 2 **STEM Participants:** 7

Major Goals: Major Goals:
In this work we investigated new ways to use concepts from geometry, probability and statistics on manifolds in order to analyze and categorize classes of data objects. This project progressed along the following three thrusts:
1. It associated classes of data objects with submanifolds and singular quotient spaces of Euclidean spaces through their adjacency matrices and (combinatorial) Laplacians. Techniques of differential geometry enabled mathematical characterization of these novel spaces and of appropriate notions of averaging of objects in these spaces.

2. Concepts from probability and statistics on manifolds were extended to adapt to the high-dimensional and geometrically complex nature of these data spaces. An appropriate probabilistic framework was developed for describing the statistical behavior of such averages.

3. From this probabilistic foundation, a variety of statistical methodologies were constructed and tested for analyzing large collections of data objects.

Accomplishments: Please see PDF file describing accomplishments.

Training Opportunities: Five PhD graduate students at Boston University received training during the reporting period. Jackson Walters and Jie (Frank) Xu are mathematics students who worked on the geometric aspects of unlabeled networks. Aleksandrina Goeva is a statistics student who worked on problems with structured and unstructured data. (Goeva has since graduated and is employed as a post-doc at the Broad Institute.) Jun Li is a statistics student who worked on the multi-network extension. (Li has since graduated and is employed at Google – Mountain View.) Nathan Josephs (not supported) is a statistics student currently working on the problem of classification with network inputs. In addition, one masters student, one PhD student, and one postdoctoral student at Notre Dame worked on this project. Respectively: (i) Maxwell Hong (not supported) worked on computational aspects of our preliminary work on latent space network modeling; (ii) Yutzu Kuo (not supported) worked on aspects of the project on classification of networks using nested Dirichlet process priors; and (iii) Kyoungjae Lee worked on a number of aspects of the project including a recent work on driving theoretical properties of a Directed Acyclic Graph model which is now accepted by Annals of Statistics with minor revision, and another paper on structure testing of a covariance graph (now under review in Journal of American Statistical Association).

RPPR Final Report
as of 31-Oct-2018

Results Dissemination: Two papers were published:

Cedric Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D. Kolaczyk (2016). Hypothesis Testing for Network Data in Functional Neuroimaging. Annals of Applied Statistics, 11(2), 725-750.

Sarpavayeva, B., Zhang, M. and Lin, L. (2018). Communication efficient parallel algorithms for optimization on manifolds. Neural Information Processing Systems 2018.

One monograph was published:
Kolaczyk, E.D. (2017). Topics at the Frontier of Statistics and Network Analysis: (Re)Visiting the Foundations. Cambridge University Press.

Five papers have been submitted:

Kolaczyk, E.D., Lin, L., Rosenberg, S., Walters, J., and Xu, J. (2018). Averages of unlabeled networks: Geometric characterization and asymptotic behavior. Annals of Statistics, (invited revision and resubmission).

Bao, D., You, K. and Lin, L. (2018). Network distance based Laplacian flow on graphs. AISTATS 2018. Submitted.

Lee, K., Lee, J. and Lin, L. (2018). Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse Cholesky factors. Annals of Statistics, (minor revision submitted).

Lee, K., Lin, L. and Dunson, D. (2018). Maximum pairwise Bayes factors for covariance structure testing. Journal of the American Statistical Association, submitted.

Lee, K., and Lin, L. (2018). Bayesian bandwidth test and selection for high-dimensional banded precision matrices. Biometrika. Submitted.

Theses/Dissertations:

Goeva, Aleksandrina. Complexity-penalized Methods for Structured and Unstructured Data. Department of Mathematics & Statistics, Boston University. May, 2017.

Li, Jun. Statistical Methods For Certain Large, Complex Data Challenges. Department of Mathematics & Statistics, Boston University. May, 2018.

Conference Proceedings

N/A

The following invited talks were given on topics relating, in whole or in part, to the work on this award:

“Statistical Analysis of Network Data.” Two-day lecture for the Bernoulli Society’s SemStat series. (With accompanying Kolaczyk (2017) Cambridge monograph.) Eindhoven, Netherlands. March, 2017.

“On the Impact of Network Inference on Network Science: Propagation of Uncertainty.” SIAM Workshop on Inferring Networks from Non-network Data (Kolaczyk, Keynote Speaker). Austin, Texas. April, 2017.

“Robust and scalable inference for big manifold data”. ‘Geometry, Statistics and Data Analysis’. RTG Statistical Sciences Symposium 2017 (Lin). May 19-20, 2017.

“Statistical Analysis of Network Data in the Context of `Big Data’: Large Networks and Many Networks.” (Kolaczyk) Department of Mathematics, Northwestern University. May, 2017.

“Nonparametric statistical inference of non-Euclidean data”. Colloquium talk (Lin). Department of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam. May 31, 2017. ?

"Statistics and Network Science: Overview and Open Problems" Annual Joint Statistical Meetings. (Kolaczyk) Baltimore, Maryland. August, 2017.

"Robust and scalable inference for big manifold data". AMS Special Session on Geometric Methods in Shape Analysis at The Ohio State University (Lin). Mar. 17-18, 2018.

"Optimization on manifolds: parallel algorithms and Bayesian optimization techniques". TGDA@OSU (Topology, Geometry, and Data Analysis @ OSU) TRIPODS workshop on Theory and Foundations of TGDA (Lin). May 21-25, 2018.

Lecturer for Summer school on "Bayesian methods for Machine Learning". Department of Mathematics and Computer Science, University of Science (Lin). Ho Chi Minh City, Vietnam. July 5-6, 2018. ?

"Intrinsic and Extrinsic Gaussian Processes on Manifolds". Eastern Asia Chapter of ISBA (Lin). Seoul, South Korea. July 12-13, 2018.

"Statistical Analysis of Network Data: Foundations (Still!) Under Construction." (Kolaczyk) Fields Institute, University of Toronto. September, 2018.

"Intrinsic and Extrinsic Gaussian Processes on Manifolds". Machine Learning Seminar series (Lin). Michigan State University. October 1, 2018.

"High-dimensional Covariance Structure Testing using Maximum Pairwise Bayes Factors". Biostatistics seminar (Lin). Northwestern University. October 29, 2018. ?

"High-dimensional Covariance Structure Testing using Maximum Pairwise Bayes Factors". Applied and Computational Mathematics Seminar (Lin). Department of Mathematics, Georgia Institute of Technology. November 5, 2018.

Honors and Awards: Eric Kolaczyk (PI) was elected a fellow of the Institute of Mathematical Statistics (IMS). (2017)

Eric Kolaczyk (PI) was elected a fellow of the American Association for the Advancement of Science (AAAS). (2018)

Lizhen Lin (Co-PI) received a NSF CAREER award. (2017).

Lizhen Lin (Co-PI) received an DARPA YFA (Young Faculty Award). (2017).

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Eric Kolaczyk
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Co PD/PI
Participant: Steven Rosenberg

Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Co PD/PI
Participant: Lizhen Lin
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Jun Li
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Aleksandrina Goeva
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Jackson Walters
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Graduate Student (research assistant)
Participant: Jie Xu
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)

RPPR Final Report
as of 31-Oct-2018

Participant: Kyoungjae Lee
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
International Collaboration:
International Travel:
National Academy Member: N
Other Collaborators:

Accomplishments Under Goals:

Primary technical accomplishments were as follows.

1. An original framework was developed for (single-layer) weighted, labeled networks. This forms the base case on which other extensions derive. It was established that the space of all such networks forms a convex manifold with corners. This in turn enabled a central limit theorem (CLT) for describing the probabilistic behavior of (nonparametric) averages of networks sampled independently and identically from a distribution supported on this space. The CLT, in turn, allowed for the derivation of hypothesis tests for comparing a collection of networks to a putative population mean, or for comparing one group of networks to another. The practical impact of this work was demonstrated through simulation and in the context of brain network data from the 1000 Functional Connectomes project. Hypothesis tests based on this new approach were found to yield substantially greater statistical power than standard approaches, effectively leveraging the advantages of multivariate testing over collective univariate testing. [This work was published in the Annals of Applied Statistics.]

2. The original framework has been extended to the case of networks that are weighted but unlabeled – a setting particularly relevant to networks in the context of privacy/security. Importantly, the lack of vertex labeling necessitates working with a quotient space modding out permutations of labels. This results in a nontrivial geometry for the space of unlabeled networks, which in turn is found to have important implications on the types of probabilistic and statistical results that may be obtained and the techniques needed to obtain them. In this work, we (i) characterize a certain notion of the space of all such networks, (ii) describe key topological and geometric properties of this space relevant to doing probability and statistics thereupon, and (iii) use these properties to establish the asymptotic behavior of a generalized (Frechet) notion of an empirical mean under sampling from a distribution supported on this space. This case has proven to be both interesting and tractable, in that while (i) strong laws of large numbers are possible for Frechet sample means fairly generally, it appears that (ii) central limit theorems result only in a more restricted setting. The development of appropriate geometric conditions necessary for a CLT to hold was fairly substantial. [This work has been revised and resubmitted at the invitation of the Annals of Statistics. Elements of this work will contribute to the dissertations of two PhD students in geometry.]

3. We have also extended the original framework of single-layer, weighted, labeled networks to multi-layer networks of the same type. Adopting the notion of graph supra-Laplacians used in this area, the core underlying geometric, probabilistic, and statistical results extend in a natural manner. However, with the introduction of network layers, there are now both a multiplicity of ways in which such a network might be interrogated (e.g., through hypothesis testing within layers, across layers, within vertex subsets, within vertex subsets across layers) and, as a result, a correspondingly daunting collection of multiple testing problems to resolve. We have adapted results in the statistics literature on high-dimensional hypothesis testing and covariance estimation to this setting. Simulations suggest that our approach to hypothesis testing and detection can provide additional benefits over the original single-layer approach (i.e., if one were to simply collapse the multi-layer network across layers) when there is structure localized in both vertex space and within only some but not all of the layers. Application of these methods to the context of multi-layer networks from computational biology, particularly in the context of personalized medicine in cancer, suggest it is possible to detect decidedly different

yet relevant biology as compared to standard tools. [To be submitted shortly to the Annals of Applied Statistics. This work formed a chapter of the dissertation of one PhD student in statistics.]

4. We have made substantial progress on extensions of classical classification frameworks to the case where inputs are network data objects. This approach incorporates Gaussian process priors and exploits connections with kernel-based learning. An appropriate notion of distance between networks lies at the heart of such approaches. We have developed the necessary methodological framework, established formal properties of statistical consistency, and implemented an initial version for application on networks of modest size (e.g., 10's to 100's of vertices). With this infrastructure now in place, we are exploring two directions (i) the implications of choice of network distance on the induced nature of the covariance kernel function and, in turn, both the theoretical and empirical performance in classification, and (ii) a comparison of our approach to standard kernel learning (i.e., SVMs). [This work is currently in manuscript form. It will be submitted upon completion and will form a chapter of the dissertation of one PhD student in statistics.]

5. Our work in unlabeled networks has opened up a variety of additional possible avenues for exploring. While our work in this direction is only preliminary at this stage, it is sufficient to indicate that success will require development of novel and nontrivial mathematical results at the intersection of geometry, networks and graphons, and statistical asymptotic theory. Given the connection of our unlabeled network problem with the increasingly prevalent use of privacy and (de)anonymization in network analysis, additional progress in this area is likely to be of particular importance. We elaborate briefly below on a handful of the interesting and challenging directions we have begun exploring to date.
    a. At present, our work has focused on the space of all networks with a fixed number of nodes. One can embed networks with N nodes into networks with N+1 nodes and to take a topological limit as N goes to infinity. It is reasonable to conjecture that laws of large numbers and CLTs hold for networks with at most N nodes for a fixed (large) N, but proving such results for all networks seems challenging. Nevertheless, the space of all unlabeled networks is an intriguing object: it contains all the information on all unlabeled networks, but it is challenging to extract particular information.
    b. There is considerable discussion about various choices of norms to measure distances between weighted networks, including l_p norms, cut distance, edit distance, and various Riemannian metrics arising in e.g. information theory. In our work, we have chosen the l_2 norm/inner product as the simplest distance measure, because from a Riemannian point of view this leads to a flat/curvature zero Euclidean metric, the easiest to handle. It is nontrivial to investigate if our results on CLTs persist for more complicated norms and Riemannian metrics. This will require techniques from both Riemannian and metric geometry.
    c. Our work has potential generalizations to ego-networks in several directions. If one node is the designated 'ego', then the quotient space of unlabeled networks with all nodes incident to that 'ego' is given by modding out an octant in R^n by a subgroup of the permutation group of the node labels. While the Euclidean metric is well-defined on the quotient, other Riemannian metrics naturally arise. As a toy example where the 'ego' node has only one incident edge, the hyperbolic metric of constant curvature -1 is a natural choice for the Riemannian metric; this metric naturally appears in information

theory as the metric on the space of normal distributions. This is certainly related to the discussion in 5(b) above.

d. As our graphs grow in size, it is natural to ask for connections between our theory and the extensive theory of graphons. There are obvious similarities, as the space of graphons is given by a function space modulo measure preserving transformations, and the space of unlabeled networks is given by a region in some R^n modulo a permutation group. However, there are also significant differences, as the function space and group of measure preserving transformations are both infinite dimensional, while the analogous spaces and groups in our setting are finite dimensional. Nevertheless, there should be some relationship between graphons and the infinite limit of networks discussed in 5(a) above.
e. The need for computationally efficient algorithms to compute sample and population means for spaces of unlabeled networks presents many challenges. Work of Jain notes that brute force algorithms are unacceptably expensive as the number of nodes increases, although he proposes short cut methods that in practice seem to converge quickly. Aside from this work, this important area seems to be wide open at present.
f. As a general method, we have noted that any compact metric space K can be isometrically embedded into the Banach space of continuous functions on K in the sup norm. We can then prove a CLT for the image of K, with the proviso that we are using the ordinary sample mean instead of the sample Frechet mean used elsewhere in our work (and the only choice of sample mean on a general metric space). Thus we have a quite general CLT, but it is not the CLT we "want." In particular, we can obtain a CLT for a compact slice of the space of unlabeled networks, but it remains to investigate the relation between these two sample means in this Banach space.